A Relational Kernel-based Approach to Scene Classification

Laura Antanas
KULeuven
Department of Computer Science
laura.antanas@cs.kuleuven.be

McElory Hoffmann
Stellenbosch University
Department of Mathematical Sciences
mr@mcelory.co.za

Paolo Frasconi
University of Florence
Department of Systems and Informatics
p-f@dsi.unifi.it

Tinne Tuytelaars
KULeuven
Department of Electrical Engineering
tinne.tuytelaars@esat.kuleuven.be

Luc De Raedt
KULeuven
Department of Computer Science
luc.deraedt@cs.kuleuven.be

Abstract

Real-world scenes involve many objects that interact with each other in complex semantic patterns. For example, a bar scene can be naturally described as having a variable number of chairs of similar size, close to each other and aligned horizontally. This high-level interpretation of a scene relies on semantically meaningful entities and is most generally described using relational representations or (hyper-) graphs. Popular in early work on syntactic and structural pattern recognition, relational representations are rarely used in computer vision due to their pure symbolic nature. Yet, today recent successes in combining them with statistical learning principles motivates us to reinvestigate their use. In this paper we show that relational techniques can also improve scene classification. More specifically, we employ a new relational language for learning with kernels, called kLog. With this language we define higher-order spatial relations among semantic objects. When applied to a particular image, they characterize a particular object arrangement and provide discriminative cues for the scene category. The kernel allows us to tractably learn from such complex features. Thus, our contribution is a principled and interpretable approach to learn from symbolic relations how to classify scenes in a statistical framework. We obtain results comparable to state-of-the-art methods on 15 Scenes and a subset of the MIT indoor dataset.

Figure 1: Sample indoor scenes belonging to classes pool inside, restaurant, bar and office (from left to right).

1. Introduction

Consider the images in Figure 1. While the pool scene can be distinguished from the others using global information and the office scene from the bar and restaurant categories using the presence of certain objects, both sources of information are prone to mistakes when differentiating between bar and restaurant scenes. In this case, in addition to the component objects, it is their complex semantic interaction that helps the scene category disambiguation. Indeed, the differentiating patterns are not the objects themselves but rather the consistent qualitative spatial and functional configurations between chairs. For example, one can describe a bar scene as having a variable number of chairs of similar size, close to each other and aligned horizontally along a counter. This high-level interpretation of a scene relies on semantically meaningful entities and is consistent across the scene category instances. It can be most generally described using relational representations [5] which can naturally capture the alignment relation among the chairs.

In this paper, we introduce a new relational representa-

tion for scene classification. The relational language builds on automatically detected semantic objects and spatial relationships that hold among them. Scenes are described as logical interpretations or, equivalently, as (hyper)-graphs. Using the language, we define qualitative spatial relations, which map object bounding boxes to higher-order relations among semantic objects. This mapping is based on domain knowledge in the form of logical rules. When applied for a particular image, the symbolic relations that hold among scene objects characterize their spatial arrangement and provide discriminative cues for the scene category. This is a more expressive representation than a fixed grid [21, 15] and more robust than continuous locations, as it allows to flexibly integrate high-level knowledge about indoor or outdoor scenes. Thus, relational representations provide a principled way to additionally represent exact metric locations as arbitrarily higher-order relations among objects. We show that such relational techniques can also improve scene classification. Moreover, our work gives a deeper insight in scene understanding by employing higher-order spatial relations among semantic objects detected using off-the-shelf object detectors. It is not only individual (relatively noisy) detections that tell the story of the scene, but also their complex dependencies.

This is typically in contrast to current trends in scene classification. They treat a scene as a whole [18], rely on independent semantic objects [25] or use scene parts without spatial information [29] or with weak spatial dependency [26, 15, 28, 27, 21, 23]. These approaches have shown that representations based on objects or parts may provide complementary information to low-level global descriptions and that the semantic configuration of the scene is important. Yet, scenes (indoor in particular) involving many objects that interact in complex semantic patterns, remain a challenge. Recent work [16, 20] has shown that stronger spatial cues in the form of geometric constraints between parts can improve results. Still, it uses a part-based model of the scene which captures only part-root dependencies, while having a fixed number of parts that do not have an explicit semantic meaning. In contrast, we consider qualitative interactions between semantic objects and explicitly describe the image as a logical interpretation.

Relational approaches were popular in early work on syntactic or structural pattern recognition [9], however today they have been rarely used to solve computer vision problems. One reason is that vision features and object detectors were not always as mature as today to support such ambitious representations. Another reason is the limitation of pure relational approaches to handle noisy data. Yet, when combined with statistical techniques, they are robust to noise [4]. Relational representations can be used in several ways to solve the scene classification problem. Related work in computer vision using such high-level representations is mostly restricted to grammars [12, 30]. We employ relational representations in a kernel-based approach that allows us to tractably learn from such complex features. In practice, we use kLog [8], a relational framework for kernel-based learning. Other related papers that make use of relations between regions of interest [11, 6, 1], or employ kernels for scene classification [24, 13, 11, 15] exist in the literature. However, none of them uses relational representations that build on semantic detected objects together with a kernel-based approach.

We evaluate our approach on the 15 Scenes dataset [15] and on 15 categories of the MIT indoor dataset [18] that are more often confused. We start from either manual annotations of the objects (when available) or automatically detected objects and show that in both cases high-level knowledge about indoor scenes can help to improve classification results. In summary, we demonstrate that in indoor environments, relations between objects are valuable if used with a flexible relational representation.

2. Scene primitives extraction

Our relational representation of a scene is built using a set of primitives. A primitive is either an object in the image with its properties, or a global property of the scene. This section describes how we obtain them from raw images. Each scene is characterized by a set of automatically detected objects in the image together with their properties.

Semantic objects Are obtained using off-the-shelf object class detectors introduced in [7, 17] which are available for use[1]. We do not use all available detectors, but we restrict to a vocabulary of 51 objects, which are more likely to appear in indoor and outdoor scenes. In addition, object classes that characterize very small objects are not considered. Although they may be discriminative for some scene categories (e.g., object `book` for category `office`), they are less likely to be accurately detected by current detectors. In practice, using 51 object detectors is reasonable, given that pretrained detectors are available. Additionally, we use less detectors than in [17]. The set of object classes considered are {`screen`, `bed`, `table`, `desk`, `counter`, `dresser`, `cupboard`, `cabinet`, `mountain`, `window`, `bookshelf`, `people`, `stair`, `door`, `railing`, `fence`, `rack`, `cloth`, `flower`, `building`, `skyscraper`, `grass`, `sky`, `tree`, `plant`, `sidewalk`, `cloud`, `tower`, `shelf`, `mast`, `ocean`, `streetlight`, `soil`, `flag`, `cue`, `pin`, `sump`, `drum`, `boat`, `bus`, `bathtub`, `bridge`, `beach`, `horse`, `cow`, `animal`, `sand`, `streetsign`, `seashore`, `truck`, `rock`}.

To increase the effectiveness of the detectors, we exploit the idea proposed in [22], where it is shown that scale-variant, or multiresolution detectors are beneficial to obtain

[1] Available at http://vision.stanford.edu/projects/objectbank/ and http://people.cs.uchicago.edu/ rbg/, respectively.

better detections. Thus, we apply the detectors at different resolutions of the image. If the size of the object to be detected is small (e.g., bottle), we run the detector at larger resolutions, otherwise at smaller ones. Also if the size of the object can vary greatly (e.g., car), a larger range of resolutions are considered. The same number of resolutions is kept (six to ten depending on the object class) across datasets. We filter out all the detections which occupy more than 70% of the image size and less than 1%. Filtered detections at one and three resolutions are shown in Figure 2(a). Next, all detections in the dataset at all resolutions are globally thresholded keeping the highest scored detections. Even after thresholding, there is a considerable number of false positive as well as missed detections. However, if a detection is a true positive, it is often obtained at many resolutions and this can be regarded as an implicit detection weight. If a detection is a false positive, its weight is typically much lower. As attributes of each object we use its class label and discretized area.

Global properties In some of our experiments, we also consider as primitives global scene properties. We use the gist [19, 20] and object bank [17], as feature descriptors to globally characterize the whole scene. Instead of directly using the raw descriptor, we use it to separately train individual classifiers on the training instances. The discrete predictions of the classifiers for each image are employed as global scene properties.

3. The relational kernel-based approach

The previous section presented the primitives that we use to describe a visual scene. Next, these are represented as a relational database and serve as input to our relational language for kernel-based learning (or kLog). Embedded in Prolog [2], the language allows to specify relational learning problems at a high-level in a declarative way. We describe next how we model our scene classification problem in kLog.

3.1. Relational scene representation

Visual scenes are represented using the classic entity/relationship (E/R) data model, a paradigm frequently used in database theory [10]. The elements of an E/R model are entity sets, corresponding to semantic objects (in Figure 3(a) they are visually depicted as rectangles), attributes that describe the objects (ovals) and relationships linking entity sets (diamonds). Entities in our application are objects with their local properties, as explained in Section 2. Spatial relationships between objects are derived from their spatial bounding boxes. The scene description is completed with global properties.

In kLog, the database scheme is directly derived from the E/R model and contains two kinds of relations: E-relations introducing entity sets and R-relations introducing relationships. In our problem , E-relations are the semantic objects. Each entity has properties and a unique identifier (underlined ovals). They can be visualized as relational facts, in Figure 2(b). The tuple $\texttt{obj}(\texttt{o3}, \texttt{bookshelf}, \texttt{large})$ specifies an object entity, where o3 is the identifier and the other attributes are, as already indicated, its class label and discretized area.

R-relations are linked to the entities that participate in the relationships. In our problem, we have spatial relationships amongst objects derived from their bounding boxes. As an example the unary relationship $\texttt{location}(\texttt{o1}, \texttt{b4})$ associates a specific object o1 with its position on a 3 by 4 rectangular grid that identifies 12 blocks in the original image; see Figure 2. In particular, for every object R and block B, $\texttt{location}(\texttt{R}, \texttt{B})$ is true iff the bounding box of R intersects B. The location of the object is conveniently specified with a relationship (and not a property) as the same object can belong to several different locations (blocks) in the image. In a relational representation, we can naturally represent sets of locations that vary in size (and thus with more discriminative power), depending on whether the object is found or not in a particular block on the grid; this is different from an attribute-based representation, where a fixed vector length is needed. We encode the global property as a special relationships of zero arity, whose attributes are associated with a scene database. For example $\texttt{global}(\texttt{office})$ is a zero arity relationship, where office represents the class predicted by a pre-trained classifier (Section 2).

A key advantage of kLog is that it supports extensional and intensional relations. Extensional relations are explicitly listed sets of facts, whereas intensional relations are defined within the language using logical rules. By deduction, as incorporated in the Prolog language, these facts can be derived from other rules or from the extensional facts. This provides a way to declaratively specify relations important to the domain. The ensemble of intensional and extensional facts describing a particular scene is called an *interpretation* and it corresponds to a small relational database [5]. Scenes are assumed to be independent.

In our scene classification application, intensional R-relations that experimentally showed to be discriminative are $\texttt{aligned\_y}/2$ (2 objects aligned on the y axis) for outdoor scenes and $\texttt{aligned\_x}/3$ (3 objects aligned on the x axis) for indoor scenes. They are derived using notions of spatial theory. As examples, $\texttt{aligned\_y}/2$ and $\texttt{aligned\_x}/3$ are defined using logical rules in the following way:

$\texttt{aligned\_y}(\texttt{O1}, \texttt{O2}) \leftarrow \texttt{obj}(\texttt{O1}, \texttt{Label1}, \_),$
$\texttt{obj}(\texttt{O2}, \texttt{Label2}, \_), \texttt{outdoor}(\texttt{O1}), \texttt{outdoor}(\texttt{O2}), \texttt{up}(\texttt{O1}, \texttt{O2}).$

$\texttt{aligned\_x}(\texttt{O1}, \texttt{O2}, \texttt{O3}) \leftarrow \texttt{obj}(\texttt{O1}, \texttt{Label1}, \_),$
$\texttt{obj}(\texttt{O2}, \texttt{Label2}, \_), \texttt{obj}(\texttt{O3}, \texttt{Label1}, \_), \texttt{indoor}(\texttt{O1}),$
$\texttt{indoor}(\texttt{O2}), \texttt{indoor}(\texttt{O3}), \texttt{right}(\texttt{O2}, \texttt{O1}), \texttt{right}(\texttt{O3}, \texttt{O2}).$

where relations $\texttt{right}(\texttt{O1}, \texttt{O2})$ and $\texttt{up}(\texttt{O1}, \texttt{O2})$ are defined

(a) Office image with object detections at 1 resolution of 400x520 (left), at 3 different resolutions (middle) and with the spatial grid on top (right).

$x = \texttt{global}(\texttt{office}), \texttt{global}(\texttt{livingroom}), \texttt{obj}(o_1, \texttt{chair}, \texttt{med}),$
$\texttt{obj}(o_2, \texttt{table}, \texttt{med})), \texttt{obj}(o_3, \texttt{bookshelf}, \texttt{large}), \texttt{obj}(o_6, \texttt{screen}, \texttt{tiny}),$
$\texttt{obj}(o_1 1, \texttt{chair}, \texttt{large}), \ldots, \texttt{location}(o_1, \texttt{b4}), \ldots, \texttt{location}(o_1, \texttt{b11}),$
$\ldots, \texttt{location}(o_{11}, \texttt{b2}), \ldots, \texttt{location}(o_{11}, \texttt{b12}), \texttt{aligned\_y}(o_3, o_1),$
$\texttt{aligned\_y}(o_3, o_5), \texttt{aligned\_y}(o_3, o_2), \texttt{aligned\_y}(o_6, o_5),$
$\texttt{aligned\_y}(o_7, o_5), \ldots, \texttt{aligned\_x}(o_4, o_3, o_8), \ldots, \}$
$y = \{\texttt{category}(\texttt{office})\}.$

(b) Logical interpretation of the image.

Figure 2: An instance in the scene classification problem; target attribute is the property, i.e., `office`, of the `category` relation.

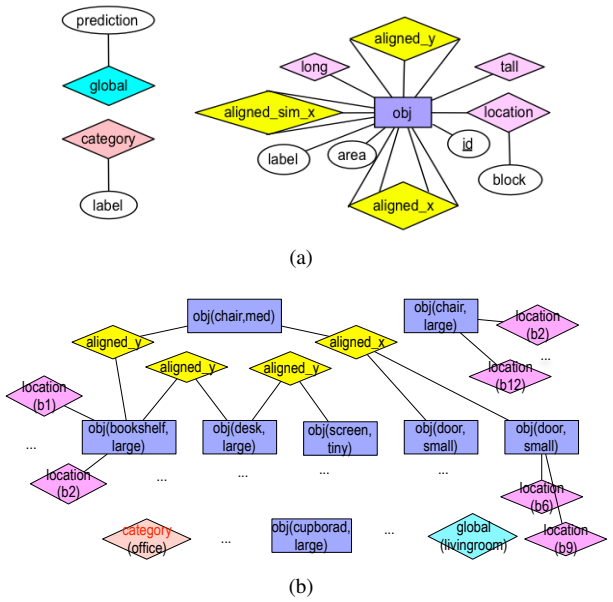

Figure 3: a) E/R modeling of the scene classification domain. Rectangles denote entity vertices, diamonds denote relationships, and circles (except obj id) denote properties. b) Graphicalized interpretation of the image.

based on the bounding boxes of the entities in a similar way. In words, `O1` is above `O2` if the minimum and the maximum y coordinates of `O1` are smaller than the minimum and the maximum y coordinates of `O2`, respectively, and if `O1` is not too much to the right or to the left (in a fuzzy way) of `O2`. The relation `indoor(O1)` specifies whether `O1` is an indoor specific object, and it helps to define the above mentioned relations only between indoor (resp. outdoor) objects.

3.2. Feature generation and learning with kLog

The learning problem at the relational level can be formalized as: given a training set of n independent interpretations $D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the goal is to learn a mapping $h : X \rightarrow Y$, where Y is the set of all scene labels. Given a new image i, and its logical interpretation (x_i, y_i) we can use h to predict y_i, that is the discrete property `label` of the target relation (i.e., `office` in Figure 2(b)). To solve it, kLog proceeds in three steps: graphicalization, feature generation and the actual learning.

Graphicalization First each interpretation x is converted into a bipartite graph G that has a vertex for each ground relation. Edges connect E-relations and R-relations: there is an undirected edge e, r if the entity identifier in e appears as an argument in r (see Figure 3(b) as an example). Vertices are annotated by grounded relations, but identifiers are removed. Role information (i.e., the position of an entity in a relationship) is retained as an edge annotation. The graph can be seen as the result of unrolling (or grounding) the E/R diagram for a particular scene. There is no loss of information associated with this step.

Feature generation Once interpretations are represented as graphs, kLog uses a graph kernel in conjunction with a statistical learner in the supervised learning setting. The kernel is a variant of the fast neighborhood subgraph pairwise distance kernel (NSPDK) [4] which: i) allows fast computations with respect to the graph size, as the graphicalization step can yield large graphs; ii) is a general purpose kernel with a flexible bias, allowing us to integrate multiple heterogeneous features. In the scene classification problem, our goal is to show the importance of logical structure.

NSPDK belongs to the large family of decomposition kernels [14] that count the number of common parts between two objects. Parts in this case are pairs of subgraphs, defined as follows. Given a graph $G = (V, E)$ and a radius $r \in \mathbb{N}$, we denote by $N_r^v(G)$ the subgraph of G rooted in v and induced by the set of vertices $V_r^v \doteq \{x \in V : d^\star(x, v) \leq r\}$, where $d^\star(x, v)$ is the shortest-path distance between x and v. For a given distance $d \in \mathbb{N}$, the *neighborhood-pair* relation is then defined as $R_{r,d} = \{(N_r^v(G), N_r^u(G), G) : d^\star(u, v) = d\}$. The kernel between two graphs is then the decomposition kernel defined by relations $R_{r,d}$ for $r = 0, \ldots, R$ and $d = 0, \ldots, D$,:

$$K(G, G') = \sum_{r=0}^{R} \sum_{d=0}^{D} \sum_{\substack{A,B\,:\,R_{r,d}(A,B,G) \\ A',B'\,:\,R_{r,d}(A',B',G')}} \kappa((A, B), (A', B')) \tag{1}$$

Although several choices are possible for κ (see [8]), in our experiments we always used the exact (hard) matching kernel where $\kappa((A, B), (A', B')) = 1$ iff (A, B) and (A', B') are pairs of isomorphic graphs. The maximum radius R and

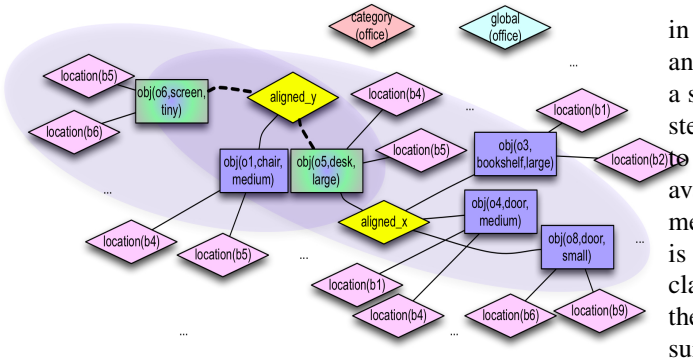

Figure 4: Illustration of NSPDK features when $D = 2, R = 2$ for a graphicalized interpretation. The sub-graph pair roots are marked in green. The path with distance $D = 2$ is marked with a dashed line and the radius as ellipses around the roots. The roots are, in this case, nodes with signature name `obj` or object entities.

the maximum distance D are kernel hyperparameters. One NSPDK feature consisting of a pair of sub-graphs is illustrated in Figure 4 for $D = 2, R = 2$.

kLog provides a flexible architecture in which only the relational language is fixed. Actual features are determined by the choice of the graph kernel but also by the definition of intensional relations, where domain knowledge can be embodied in the form of rules. In this setting, experimenting with alternative feature spaces is rapid and intuitive. Finally, any statistical learner can be used to learn from the obtained feature vectors.

4. Experiments

We perform experiments on a subset of 15 categories of the MIT indoor dataset (15MIT) introduced in [18] and the 15Scenes in [15]. The first dataset contains many indoor scene categories and poses a challenging classification problem. We consider fifteen categories that are more likely to be confused mostly due to the fact that the intra-class variability of such categories is high and therefore, the average category image is a uniform field. The categories considered are: {`auditorium`, `bedroom`, `computer room`, `classroom`, `restaurant`, `waiting room`, `inside bus`, `bar`, `office`, `fast − food restaurant`, `concert hall`, `living room`, `dining room`, `meeting room`, `office`, `kitchen`}.
The second dataset contains six indoor categories {`kitchen`, `bedroom`, `living room`, `meeting room`, `office`, `store`}, and nine outdoor categories {`suburb`, `tall building`, `inside city`, `industrial`, `highway`, `coast`, `street`, `open country`, `forest`, `mountain`}. Our goal is to show that, especially on such difficult problems, combining semantic features and high-level, rich relations between them, improves classification.

For evaluation we use the same training/test split as in [18, 15], where each scene category has about 80 training and 20 test images. For the MIT dataset we do, in addition, a set of experiments in which we use only ground truth instead of object detections in the image. Since we also want to acquire the impact on the results when rich information is available, it is also this setting that we employ in our experiments. As the data for training pairwise category classifiers is balanced, we report the more appropriate overall multiclass prediction accuracy as evaluation measure, instead of the average multiclass prediction accuracy (which is more suitable when the data is unbalanced).

4.1. Features used

We experiment with the following features:
· zero arity relationship `global`: captures i) the gist of a scene using spectral and coarsely localized information across the image [18], and ii) a collection of scale invariant response maps of a large number of pre-trained object detectors [17]. We integrate them simultaneously in a discrete way, i.e., as the output of separate classifiers.
· object local properties: i) object entity is assigned as attribute the generic category `object`, such that no identity of the detected object is assumed; ii) object entity is assigned as attribute the object category label and its discretized area.
· unary relationship `location`: captures the position of the object on a grid that identifies 12 blocks in the image.
· unary relationships `tall/long`: if an object bounding box is taller (longer) than 2/3 of the image height (width).
· higher-order relationships defined on the y axis: `aligned_y2` captures the alignment on the y axis of two objects bounding boxes; `aligned_sim_y2` imposes, in addition to the vertical alignment, that the objects must be similar in appearance, i.e., they have the same class labels;`aligned_sim_y3` is similar to `aligned_sim_y2` but defined between three objects.
· higher-order relationships defined on the x axis: `aligned_x3` captures the alignment on the x axis of three objects; `aligned_sim_x3` imposes, in addition, that the objects have the same class labels; `aligned_x4` is a quadruple relation that holds between four objects that have similar, but interleaved (e.g., `A − B − A − B`) object class labels.

In all defined relations, except the unary ones, the objects involved in the relation represent at the same time either indoor, outdoor or natural objects. This condition follows naturally as we do not want to add, for example, a relation between a car and a desk. We refer to all relationships as `all relations` and to local attributes and all relationships, except global properties, as `all`.

4.2. Evaluation and results

The purpose of the paper is to answer the following questions: **1)** Do symbolic relations between objects improve

scene classification? **2)** How does the quality of the detections influence the classification results? **3)** Do relational representations provide complementary information to global descriptors?

To answer question 1), we analyze the impact of the symbolic information gradually by combining different features. We incrementally incorporate richer and richer relational information to assess the importance of the features used. As a baseline, we use the generic discrete label object as the class of the objects. Next, we replace the object label with the available class label for each detection to see the importance of the semantic features. We add discretized area information (in eight intervals) on the objects and then add gradually functional and spatial information by incorporating the user defined symbolic relations: `location`, `tall/long` and more complex relations as listed in Section 4.1. We report performance results in Table 1.

Adding the `area` attribute, unary relations `location` and `tall/long` , separately, improves classification results on both 15MIT and 15Scenes. When they are combined, classification performance increases. We get more improvement when the rest of the relations are added. The justification of this result is that when relations between objects are injected in the graph as information about their configuration in the scene, the feature mapping encodes even more high-level, discriminative information about the scene. This increases the classification performance. The hyperparameters of the kernel that gave the best results were R=0 and D=0 for `label` + `area`. In words, this is equivalent to counting object interest points (or roots with signature name `obj`) on the graph. For the cases when relations were also used, the best performance was obtained when R=1 (R=2) and D=0. This is equivalent to counting on the graph (pairs of) object interest points with a same spatial constraint. Thus, including rich relations and using a larger radius improve classification results, which lets us conclude that indeed symbolic relations are helpful discriminative features.

To answer question 2) we replace object detections with the manually annotated objects and their bounding boxes (15MIT annot). Our goal is to show that also when starting from less noisy and rich detectors, relational representations can improve indoor scene classification (see Table 1). Again, we gradually include high-level information. We notice that the weaker the local information is, the more relations help. This can be interpreted as when the local semantic information is strong, the impact of qualitative relations is more limited. However, we get an improvement of 1.5% on the annotations with the help of relations. Out of the 15 categories considered, only 7 had annotations also on the test set and not all instances in the 7 categories were annotated. This leads to typically a lower accuracy for 15MIT annot than for 15MIT, since all unannotated instances were classified as belonging to the same class.

Features		Overall Accuracy (%) 15MIT annot	15 MIT	15 Scenes
L	generic object	3.3	4.0	6.7
	labels	33.1	28.4	45.8
	labels+area	35.5	39.5	64.4
L+R	labels+loc	35.5	34.1	66.7
	labels+tall/long	34.8	35.1	58.7
	labels+area+loc	36.2	46.0	67.6
	labels+area+loc+tall/long	36.5	46.6	68.3
	labels+area+all relations	37.2	47.5	69.7
G	gist [23, 20]	36.6	36.6	52.3
	object bank [17]	49.7	49.7	80.9
G+L +R	global+all	60.5	54.2	82.0

Table 1: Overall accuracy for the considered datasets. L denotes local object attributes, R denotes unary/binary/ternary/quadruple relationships and G denotes global information.

We answer question 3) by combining global properties `global`[2] with all features. Their combination gave the best results, which answers the question affirmatively. Their complementarity is qualitatively illustrated in Figure 5. The bottom row shows mistakes made by our relational approach. These are mainly due to noisy detections where discriminative qualitative relations between meaningful detections could not be established or where relations between chairs were not properly captured by our spatial theory rules in Prolog. Qualitative relations helped in several cases (top row) by capturing configurations between meaningful detected objects at different resolutions (e.g., relations between chairs in the meeting room in the third figure from the left). We note that the complementarity is best visible for 15MIT annot, where the object detections and thus, the relations, are more precise.

Many alternative statistical learners can be used on the feature vectors. We used a standard implementation of support vector machines with one-vs-one handling of multiclass classification [3], which is integrated via a wrapper in kLog. We choose the cost of the SVM and the set of discriminative relations by performing internal 10 fold cross-validation on the training set.

5. Conclusions

The aim of this paper was to show that relational representations are beneficial and can also improve scene classification results. To this end, we use global properties, semantic object detections with their local properties, unary relations on the objects and complex spatial relations among

[2]The sole gist/object bank give the same result for 15MIT annot and 15MIT as they do not use any object bounding box annotations explicitly.

Figure 5: Images missclassified by object bank/gist and correctly classified by our relational approach (top). Images where our approach fails (bottom).

them to build a relational database of scenes. This relational database forms the input to the statistical relational framework, kLog, which models the scene classification problem in a principled and declarative way. We obtain results competitive with state-of-the-art.

Laura Antanas is supported by the European Commission under contract FP7-248258-First-MM.

References

[1] M. R. Boutell, J. Luo, and C. M. Brown. Scene parsing using region-based generative models. *IEEE Transactions on Multimedia*, 9(1):136–146, 2007. 2

[2] I. Bratko. *PROLOG Programming for Artificial Intelligence*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition, 1990. 3

[3] C. Chang and C. Lin. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27, 2011. 6

[4] F. Costa and K. D. Grave. Fast neighborhood subgraph pairwise distance kernel. In *ICML*, pages 255–262, 2010. 2, 4

[5] L. De Raedt. *Logical and Relational Learning*. Springer, 2008. 1, 3

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, pages 229–236. Ieee, 2009. 2

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627 –1645, sept. 2010. 2

[8] P. Frasconi, F. Costa, L. D. Raedt, and K. D. Grave. klog: A language for logical and relational learning with kernels. *CoRR*, abs/1205.3981, 2012. 2, 4

[9] K. Fu. *Syntactic methods in pattern recognition*, volume 112. Elsevier Science, 1974. 2

[10] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall Press, Upper Saddle River, NJ, USA, 2 edition, 2008. 3

[11] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *In ICCV*, pages 1458–1465, 2005. 2

[12] F. Han and S. Zhu. Bottom-up/top-down image parsing with attribute grammar. *TPAMI*, 31(1):59–73, 2009. 2

[13] Z. Harchaoui. Image classification with segmentation graph kernels. In *CVPR*, pages –1–1, 2007. 2

[14] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California at Santa Cruz, 1999. 4

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2 of *CVPR '06*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. 2, 5

[16] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, pages 2735 – 2742, june 2012. 2

[17] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, Vancouver, Canada, December 2010. 2, 3, 5, 6

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42:145–175, May 2001. 2, 5

[19] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006. 3

[20] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 2, 3, 6

[21] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb. Reconfigurable models for scene recognition. In *CVPR*, pages 2775 –2782, june 2012. 2

[22] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, pages 241–254, Berlin, Heidelberg, 2010. Springer-Verlag. 2

[23] A. Quattoni and A. Torralba. Recognizing indoor scenes. *CVPR*, 0:413–420, 2009. 2, 6

[24] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The nbnn kernel. In *ICCV*, pages 1824–1831, 2011. 2

[25] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007. 2

[26] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *PAMI*, 33(8):1489–1501, 2011. 2

[27] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, pages 9 –16, june 2010. 2

[28] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang. Hierarchical gaussianization for image classification. In *ICCV*, pages 1971–1977, 2009. 2

[29] J. Zhu, L.-J. Li, F.-F. Li, and E. P. Xing. Large margin learning of upstream scene understanding models. In *NIPS*, pages 2586–2594, 2010. 2

[30] L. Zhu, Y. Chen, Y. Lin, C. Lin, and A. Yuille. Recursive segmentation and recognition templates for image parsing. *TPAMI*, 34(2):359–371, 2012. 2